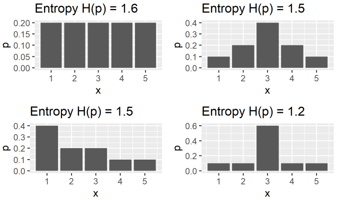
Introduction to Machine Learning

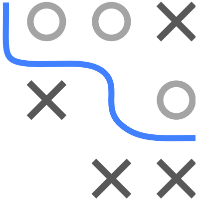
Entropy II



Learning goals

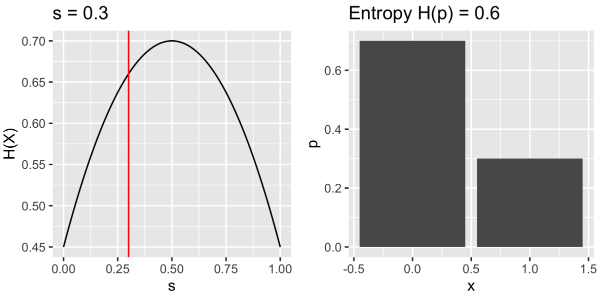
- Further propterties of entropy and joint entropy
- Understand that uniqueness theorem justifies choice of entropy formula
- Maximum entropy principle

ENTROPY OF BERNOULLI DISTRIBUTION



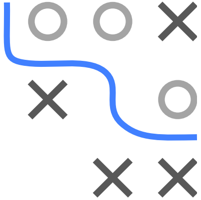
Let X be Bernoulli / a coin with $\mathbb{P}(X = 1) = s$ and $\mathbb{P}(X = 0) = 1 - s$.

$$H(X) = -s \cdot \log_2(s) - (1 - s) \cdot \log_2(1 - s).$$



We note: If the coin is deterministic, so $s = 1$ or $s = 0$, then $H(s) = 0$; $H(s)$ is maximal for $s = 0.5$, a fair coin. $H(s)$ increases monotonically the closer we get to $s = 0.5$. This all seems plausible.

JOINT ENTROPY



- The **joint entropy** of two discrete random variables X and Y is:

$$H(X, Y) = H(p_{X,Y}) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2(p(x, y))$$

- Intuitively, the joint entropy is a measure of the total uncertainty in the two variables X and Y. In other words, it is simply the entropy of the joint distribution $p(x, y)$.
- There is nothing really new in this definition because $H(X, Y)$ can be considered to be a single vector-valued random variable.
- More generally:

$$H(X_1, X_2, \ldots, X_n) = -\sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} p(x_1, x_2, \ldots, x_n) \log_2(p(x_1, x_2, \ldots, x_n))$$

ENTROPY IS ADDITIVE UNDER INDEPENDENCE

7. Entropy is additive for independent RVs.

Let X and Y be two independent RVs. Then:

$$
\begin{aligned}
H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2(p(x, y)) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_X(x) p_Y(y) \log_2(p_X(x) p_Y(y)) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_X(x) p_Y(y) \log_2(p_X(x)) + p_X(x) p_Y(y) \log_2(p_Y(y)) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_X(x) p_Y(y) \log_2(p_X(x)) - \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_X(x) p_Y(y) \log_2(p_Y(y)) \\
&= -\sum_{x \in \mathcal{X}} p_X(x) \log_2(p_X(x)) - \sum_{y \in \mathcal{Y}} p_Y(y) \log_2(p_Y(y)) = H(X) + H(Y)
\end{aligned}
$$

THE UNIQUENESS THEOREM ▸ KHINCHIN, 1957

Khinchin (1957) showed that the only family of functions satisfying

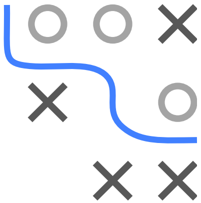
- $H(p)$ is continuous in probabilities $p(x)$
- adding or removing an event with $p(x) = 0$ does not change it
- is additive for independent RVs
- is maximal for a uniform distribution.

is of the following form:

$$H(p) = -\lambda \sum_{x \in \mathcal{X}} p(x) \log p(x)$$

where λ is a positive constant. Setting $\lambda = 1$ and using the binary logarithm gives us the Shannon entropy.

THE MAXIMUM ENTROPY PRINCIPLE ▸ JAYNES, 2003



Assume we know M properties about a discrete distribution $p(x)$, given as moment conditions for functions $g_m(\cdot)$ and scalars α_m:

$$\mathbb{E}[g_m(X)] = \sum_{x \in \mathcal{X}} g_m(x)p(x) = \alpha_m \text{ for } m = 0, \ldots, M$$

- Principle of maximum entropy: Among all distributions satisfying these constraints, choose the one with maximum entropy
- Intuitively, this ensures that amount of prior assumptions on $p(x)$ are minimal (avoids “overfitting”)
- We already saw an application of this: for the (trivial) constraint $\sum_{x \in \mathcal{X}} p(x) = 1$ ($g_0(x) = 1 = \alpha_0$), we derived the uniform distribution as having maximum entropy

Maxent distribution given M constraints can be computed from Lagrangian with multipliers $\lambda_1, \ldots, \lambda_M$. Finding the optimal λ_m means finding the constrained maxent distribution.

THE MAXIMUM ENTROPY PRINCIPLE

The Lagrangian for this problem using base e is given by:

$$L(p(x), (\lambda_m)_{m=0}^{M}) = -\sum_{x\in\mathcal{X}} p(x)\log(p(x)) + \lambda_0\Big(\sum_{x\in\mathcal{X}} p(x) - 1\Big) + \sum_{m=1}^{M}\lambda_m\Big(\sum_{x\in\mathcal{X}} g_m(x)p(x) - \alpha_m\Big)$$

Finding critical points $p^*(x)$:

$$\frac{\partial L}{\partial p(x)} = -\log(p(x)) - 1 + \lambda_0 + \sum_{m=1}^{M}\lambda_m g_m(x) \overset{!}{=} 0 \iff p^*(x) = \exp(\lambda_0 - 1)\exp\Big(\sum_{m=1}^{M}\lambda_m g_m(x)\Big)$$

This is a maximum as $-1/p(x) < 0$. Since probs must sum to 1 we get

$$1 = \sum_{x\in\mathcal{X}} p^*(x) = \frac{1}{\exp(1-\lambda_0)}\sum_{x\in\mathcal{X}}\exp\Big(\sum_{m=1}^{M}\lambda_m g_m(x)\Big) \Rightarrow \exp(1-\lambda_0) = \sum_{x\in\mathcal{X}}\exp\Big(\sum_{m=1}^{M}\lambda_m g_m(x)\Big)$$

Plugging $\exp(1-\lambda_0)$ into $p^*(x)$ we obtain the constrained maxent distribution:

$$p^*(x) = \frac{\exp\sum_{m=1}^{M}\lambda_m g_m(x)}{\sum_{x\in\mathcal{X}}\exp\sum_{m=1}^{M}\lambda_m g_m(x)}$$